Practice of Disclosure Limitation Techniques

by

Lu Shen

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Master of Science.

Baltimore, Maryland

August, 2014

Abstract

In recent years, there has been increasing concern about the identification of personal or corporate information in published data; the privacy issue. Much of this interest has been focused on issues relating to release of micro (e.g., patient-specific) and tabulated data. Increasingly, people ask: "Does de-identification work or not?" In this project we review a subset of the current disclosure limitation methods, including suppression and aggregation, swapping, random noise and synthetic data. Then, we assess the performance and disclosure risk associated with a few of the methods. To accomplish this, we use the microdata collected by the PREMIER Collaborative Research Group (Elmer et al., 2006). Specifically, random noise and synthetic data methods are evaluated by comparing the results obtained from the modified data with those obtained from micro data. Furthermore, we compare the modified data in regards to disclosure risk using alpha/beta measures and differential privacy method.

Acknowledgments

First and foremost, I would like to express my sincerest and warmest gratitude to my advisor Dr. Thomas Louis for his supreme patience, exceptional support, and great effort throughout my thesis writing. His teaching and guidance helped me in all the time of research and writing of thesis. This thesis would not have been completed without his immense knowledge and encouragement.

I am grateful to Dr. Elizabeth A. Stuart who guided me as my second advisor, for her constructive comments and valuable suggestions. Additionally, I would like to thank the faculty in the department of Biostatistics for their support and help during my study at Johns Hopkins Bloomberg School of Public Health.

I would like to thank my best friend Dr. Hui Lin from Iowa State University, who helped me carefully reading and editing during the writing of thesis. I would also thank Dr. Hui Zhou from Columbia University for his kind, constant help and support. Also, thanks to Stephen Cristiano for your advice and invaluable friendship.

I would like to give my special thanks to my partner Xiyu Zhou. Your love, patience, and care enabled me to complete this thesis.

ACKNOWLEDGMENTS

Finally to my parents, Lilin Shen and Fuming Liu, receive my deepest gratitude and love for their dedication and many years of support during my undergraduate and graduate studies that provided the foundation for this work.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The U.S. Census Bureau collects its survey and census data under Title 13 of the U.S. Code. This provides a legal barrier to the Census Bureau from releasing any data "...whereby the data furnished by any particular establishment or individual under this title can be identified." Statistical agencies have made great efforts to maintain the level of confidentiality while releasing the data for public use. Especially for data released by the U.S. Census Bureau, researchers in government applied different types of data disclosure avoidance techniques in order to meet their legal obligations, reduce disclosure risk, but at the same time release high quality, informative data to the public.

As discussed by (Louis, 2013), there are several ways to make data available while protecting confidentiality. Successes include a variety of techniques used to produce the Local Employment and Housing Dynamic (LEHD) website, the American fact

finder, "On the Map," and the micro-data analysis system (MAS). In these and other contexts, full micro-data provide complete information (i.e., the highest quality data), but releasing it would result in 100% disclosure risk, and vice-versa in that releasing no data provides no information with 0% disclosure risk. The goal of disclosure avoidance or limitation is to achieve an acceptable trade off: under a certain disclosure risk (society's decision), the goal is to optimize informativeness of the data. In the current "big data" context with active intruder threats it's not an easy goal to accomplish. The trade-off is very similar to that for an ROC curve related to diagnostic testing, with sensitivity and (1 - specificity) replaced by information and disclosure risk.

In this report, we first review the current methods of disclosure risk and access the disclosure risk and information loss. Later, we apply methods to "micro-data" and compare results.

Chapter 2

Methods

2.1 Overview of disclosure limitation techniques

Over the past 20 years many statistical and computer science approaches have been developed and implemented to meet the goal of limiting disclosure risk. The main methods that have been used include cell suppression, data swapping, noise masking and generation of fully or partially synthetic data. We briefly review each.

Generally speaking, cell suppression limits disclosure by complete deletion of information. The data swapping method involves repeatedly moving pairs of columns or rows around while maintain higher-level totals. For instance, we can randomly swap 30% of race or of martial status information. Both cell suppression and data swapping preserve marginal totals. For the noise masking method, depending on the

type of data, we either multiply or add noise. Increasing the noise increases protection. There is another similar method called synthetic data generation. It can be summarized by "don't keep everything." First, we build a prediction model with all or a subset of the complex relationships in the original, micro-data. Then, we generate a random data set using that model and release the random generated data to the public. If we do a regression using both the generated data set and actual data set, the regression slope we get from each should be similar. One disadvantage is that it's hard to capture all the relationship in the actual data when we use to build the prediction model. Especially when we have too any variables in a data set, it's also time consuming to figure out all the relationship in the actual data set. It'll be more problematic if we fail to include one specific relationship to generate synthetic data. Suppose we fail to include the relationship between age and income in the prediction model to generate synthetic data. People who use the resulting synthetic data to do the analysis will get the conclusion that there's no relationship between income and age. We can use PCA method to look at possibly correlated variables to reduce variables or regression analysis to look at the relationship between variables when we try to build the prediction model. Valuable information will be lost if we fail to use variable of interest to do the prediction model. For example, if the prediction model is conditional on income bands of $10,000 and the analyst is interested to use income bands of size $5,000. One example is in a two-way contingency table, if we worry about letting out the inside values, we can maintain the row or column totals

or percentages and use the generated values fill out the inside. We need to make sure that the inside values sum to the row or column totals or produce the marginal percentages.

2.1.1 Aggregation

Aggregation is a method that turns atypical records into typical records. For example, there may be only one or two people with a particular combination of house keys in a county, but there may be many people with such keys in a state. Releasing this type of information for a county might put these one or two people at disclosure risk (with house keys, entry risk!) since they will be easy to find. Aggregating to the state level provides protection, but reduces information. Similarly, data aggregation increases disclosure protection, but reduces informativeness.

Another example might be to report exact values only below specified thresholds. For example, we could report age in 5 or 10 year intervals rather than reporting exact ages e.g., report people with age 23 as age 20-25, or recode income in a $5000 range or even do county-specific but not census tract-specific data. These recoding process will generally affect the detailed summary statistics for the full data (Karr and Reiter, 2014). The most problematic issue is about uncongeniality. "Uncongeniality essentially means that the analysis procedure does not correspond to the imputation model. The uncongeniality arises when the analyst and the data generating person have access to different amounts of information and have different assessments."

(Meng, 1994) U.S. Census always solve this problem by asking the analyst what type of analysis they want to perform up front before they provide the analyst with the generated data set. In this way, we can make sure that the generated data set will contain the specific relationship and variable of interest that meet the analysts' needs.

2.1.2 The data swapping method

Discussed by (Fienberg et al., 1996), suppose X is an $n \times p$ matrix. The matrix masking of micro-data X provides the user with transformed data $Z = AXB + C$, where A transforms cases, B transforms variables, and C blurs the entries of AXB. The well-known approaches are as follows:

- Release a subset of data (delete rows of X)
- Include simulated data (add rows to X)
- Add random perturbations to X
- Exclude attributes (delete columns of X)
- Release the variance-covariance matrix (choose $A = X^T$)

The deletion of rows and release the subset of data is often called as cell suppression method. Cell suppression is widely used for data on establishments since some critical variables can be exploited to disclosure everything about a unit in the dataset. Some critical variables may help intruder to identify a respondent. The disadvantage for using this method is that we may lose some valuable information, however, we protect the respondents.

(Dalenius and Reiss, 1978) was the first to discuss the data swapping method as a way of protecting data that contain categorical variables. Table 2.1 contains three variables for five respondents. Suppose that age is a sensitive variable, then we can swap Age variables on any two pairs of records. On the left is the original data set and the resulting swapped data set is on the right. This method proceeds done by only swapping one variable. Because Table 2.1 is small, it's possible to identify the respondent by trial and error. However, it's hard to identify them in a large $n \times p$ data set.

Table 2.4 illustrates data swapping in a three-way contingency table. The observation for (1,2,1) cell is moved to the other layer (i.e., into the (1, 2, 2) cell). The observation from the (3,1,2) cell is moved to the first layer (i.e., the (3,1,1) cell). Therefore, moving from the original table (on the left) to the swapped pair observations table (on the right), we end up preserving the two-way totals, n_{ij+} and the one way total, n_{++k}. This procedure can help to preserve the marginal total but there is always some drawbacks. One drawback is when we use this method to swap categorical variables, we need to be careful that we may decrease the usefulness of data by creating some strange combinations. Some strange combinations may not make much sense. Another drawback is that it may greatly affect the data's analytical value and distort the correlations between variables.

2.1.3 Noise Perturbation

Following notation from (Nayak et al., 2010), let Y with values $Y = (y_1, y_2, \ldots, y_n)$ be the actual data for n units and $Z = (z_1, z_2, \ldots, z_n)$ be the masked data. Let μ_y and σ_Y^2 be the mean and variance of Y. The basic mechanism for random noise perturbation is to generate n numbers $(r_1, r_2, ..., r_n)$ from a given noise distribution independently and apply them to the y values, either by addition or multiplication (i.e., the masked data set Z, where $z_i = y_i + r_i$ or $z_i = y_i \times r_i$). Usually, the data agencies use mean 1 for noise multiplication and mean 0 for noise addition.

For noise multiplication $Z = Y \times R$, it's clear that

$$E[Z] = \mu_Y \tag{2.1}$$

$$\begin{aligned} V[Z] &= V[E(Z|Y)] + E[V(Z|Y)] \\ &= \sigma_Y^2 + \sigma_R^2[\sigma_Y^2 + \mu_Y^2] \\ &= (1 + \sigma_R^2)\sigma_Y^2 + \mu_Y^2\sigma_R^2 \end{aligned} \tag{2.2}$$

As shown in equations 2.1 and 2.2, the mean $\bar{Z}$ of the masked data set Z is an unbiased estimator of μ_Y, but the variance overestimates σ_Y^2.

Noise multiplication could also apply for more than one variable when you generate the noise perturbation factors independently. For each variable, the noise perturbation factors could come from different distribution. For example, in addition to Y variable, we have another variable W in the raw data set. The masked data set contains unchanged variable W and Z. Following from the equation 2.2, we then

have

$$\rho(Z,W) = \left[\frac{\sigma_Y^2}{(1+\sigma_R^2)\sigma_Y^2 + \mu_Y^2\sigma_R^2}\right]^{1/2} \rho(Y,W) \tag{2.3}$$

where $\rho(Z,W)$ denotes the correlation between Z and W.

Hence, noise perturbation (addition or multiplicity) will attenuate correlations as shown in equation 2.3. In order to account for this problem, we can generate multivariate noise as alternative solution.

$$\begin{aligned} E(Z_1Z_2) &= E(Y_1R_1Y_2R_2) = E(Y_1Y_2R_1R_2) = E(Y_1Y_2)E(R_1R_2) \\ &= E(Y_1Y_2)E(R_1)E(R_2) = E(Y_1Y_2) \end{aligned} \tag{2.4}$$

$$\begin{aligned} cov(Z_1,Z_2) &= E(Z_1Z_2) - E(Z_1)E(Z2) = E(Y_1Y_2) - E(Y_1)E(Y_2) \\ &= cov(Y_1,Y_2) \end{aligned} \tag{2.5}$$

As shown in 2.5, independent noise multiplication with mean 1 does not bias the sample covariance.

2.1.4 The Synthetic Data Method

The synthetic data method can be implemented in different ways. One way is called "full synthetic" by synthesizing all variables for all records. Another way is called "partial synthesis" by synthesizing a subset of variables or a subset of records. Partially synthetic data allows users to select custom geographies in "On The Map" `http://onthemap.ces.census.gov/`. For both fully synthetic data and partially synthetic data, the idea is to create a statistical model and then generate one or more

pseudo-data sets from it. Specifically, synthetic data are produced by fitting a model to X and generating data from it's posterior predictive distribution $Z \sim model(X)$.

The most important thing for synthetic method is that we must include the association of interest when fitting the model. If we fail to include the association of interest, the user will get a estimate that is statistically close to 0 for the association using the pseudo-data. (Rubin, 1993) discussed full simulation based on multiple imputation. Generally, it first randomly samples units from the sampling frame for each synthetic data set, and then unknown data values for units in the synthetic samples are imputed. The biggest problem for this method is to consider the relationship within the data set (i.e., the structure within the data set).

For partially synthetic data, as discussed by (Reiter, 2003), let $I_j = 1$ if unit j is selected in the original survey, and $I_j = 0$ otherwise. Then we have $I = (I_1, ..., I_N)$. Let Y_{obs} be $n \times p$ matrix of actual survey data for units with $I_j = 1$. Let X be the $N \times d$ matrix of design variables for all N units in the population. The observed data set is therefore $D = (X, Y_{obs}, I)$. First, we randomly select values from the observed data set and replaced with imputations. Second, impute new values to replace those selected values. Let $Z_i = 1$ if unit j is selected to be replaced with synthetic values and $Z_i = 0$ for data with unchanged values and $Z = (Z_1, ..., Z_n)$. Let $Y_{rep,i}$ be all the imputed values in the i^{th} synthetic data set and let Y_{nrep} be all the unchanged values of Y_{obs}. $Y_{rep,i}$ are generated from the Bayesian posterior predictive distribution of $(Y_{rep,i}|D, Z)$ and Y_{nrep} are the same in all synthetic data sets.

Therefore, each synthetic data set, (d_i) contains $(X, Y_{rep,i}, Y_{nrepe}, I, Z)$. Imputations are made independently for $i = 1, ..., m$ times for m different synthetic data sets.

2.2 Disclosure Risk

2.2.1 Measuring Risk

Discussed by (Reiter, 2003), measuring disclosure risks can be assessed several ways; here are two: (1) estimating the number of records released in the sample whose characteristics are unique in the population. (2) estimating the probabilities that records can be identified from the released data. As proposed by many authors, measuring disclosure risk depends on uniqueness; unique units have a higher risk than non-unique units. However, uniqueness can not be assessed easily.

Many authors discuss measuring disclosure risk using a Bayesian approach. Parameters are considered as coming form a prior distribution and the prior is updated in the light of the data, producing a posterior distribution and then generating all inferences based on this posterior distribution. Generally, disclosure risk associated with the prior is measured by two rules: (1) the $n \sim k$ dominance rule, in which a sensitive cell is defined when the sum of the contributions of n or fewer respondents represents more than a fraction k of the total cell value; (2) p% rule, in which protect the largest company values in a given cell from upper estimation to within p%. In general, many literatures state that the p% rule is preferred to the $n \sim k$ dominance rule since it provides more suppressions of the cell than $p \sim q$ rule (Hundepool et al.,

2012). To help better understanding p% rule, let T be the total value of a specific cell, L be the value of the largest contributor to the cell, S be the value eof the second largest contributor to the cell and p be required percentage of protection. The p% rule states that a cell must be suppressed if $T - L - S < (p/100) \times L$.

(Oganian and Domingo-Ferrer, 2003) proposed using Shannon's entropy of relative contributions to a table cell approach,

$$H(X) = \sum_{i=1}^{n} \left(\frac{x_i}{x}\right) \log_2 \left(\frac{x_i}{x}\right) \tag{2.6}$$

where the x_i are contributions to cell i and $x = x_1 + x_2 + ... + x_N$. A cell is considered sensitive when with $t \in [0, 1]$,

$$\frac{H(X)}{\log_2(N)} < t \tag{2.7}$$

They also suggest using the reciprocal of conditional entropy as a posterior measure for disclosure risk,

$$\begin{aligned} DR(X) &= \frac{1}{H(X|Y=y)} \\ &= \frac{1}{-\sum_x p\left(\frac{x}{y}\right) \log_2 p\left(\frac{x}{y}\right)} \end{aligned} \tag{2.8}$$

where X is the original cell variable, and Y is a variable representing the knowledge of the actual data (equals some specific value y). Under Formula 2.8 the more the uncertainty about the value of original cell of X, the less the disclosure risk, and vice versa. In order to compute Formula 2.8, we need to find the set $S_y(X)$ of possible values of X given the constraints y, and estimate the probability of the cell $X = x$ conditional on $Y = y$.

Using the uniform distribution, Formula 2.8 simplifies to,

$$DR_{unif}(X) = \frac{1}{\log_2\{m(S_y(X))\}} \tag{2.9}$$

where $m(S_y(X))$ is the number of cell values in $S_y(X)$.

2.2.2 Measures of perturbation

As discussed by (Massell and Funk, 2007a), the desired amount of perturbation always depends on the sensitivity status of the cell. As previously discussed, we use the standard p% rule to determine the sensitivity of the cell. A cell is considered sensitive if on the basis of its value it is possible to estimate the contribution of an individual respondent to that cell to within p percent of the value of its contribution. Under the noise perturbation method, it's possible that we can't meet these goals and must tolerate some under-perturbed and over-perturbed cells. It's desirable to have have a general formula to measure the amount of under-perturbation of sensitive cells and over-perturbation of safe cells. Generally, this can be done using a modeling approach and simulations. The modeling approach takes long time to do in each instance and so is not realistic. Simulations are usually performed by trial and error analysis to find acceptable parameters and is often termed "calibration of the noise distribution."

Formula 2.10–2.11 shows a global scalar measure of under-perturbation and over-perturbation. It's performed by two types of measures: alpha measures and beta measures. Alpha measures under-perturbation of sensitive cells; Beta measures over-

perturbation of both safe and sensitive cells. Also, you can not minimize alpha measure and beta measure at the same time. Using Formula 2.10–2.11, we can calculate alpha measures and beta measures for a given table.

Alpha: A Global Scalar Measure of Under-Perturbation

$$\alpha = \frac{\sum_{\text{all sensitive}} \max\left(0, 1 - \left(\frac{\text{Actual perturbation}}{\text{Nominal perturbation}}\right)\right)}{\#\text{ Sensitive}} \tag{2.10}$$

Beta: A Global Scalar Measure of Over-Perturbation

$$\beta = \frac{\sum_{\text{all safe}} \left|\frac{X-Y}{X}\right| + \sum_{\text{all protected sensitive}} \left|\frac{|X-Y| - \text{prot}}{X}\right|}{\#\text{ Safe} + \#\text{ Protected Sensitive}} \tag{2.11}$$

In order to decide when a particular level of noise is too low or too high, a great number of computational experiments with a variety of noise distributions are necessary to determine acceptably small values for the alpha and beta measures.

Protection Multiplier (PM) is another distributional measure of Under-Perturbation,

$$\text{PM} = \frac{|\text{Perturbation from Noise}|}{\text{Suggested Perturbation}} \tag{2.12}$$

(Massell and Funk, 2007a) have found that some distributional measures of under and over perturbation to be more useful than the alpha measures and beta measures. They also pointed out the only disadvantage is that a density must be produced, either in the form of a table or graphically. Use Formula 2.12 to compute PM for each sensitive cell. See (Massell and Funk, 2007b) for more details in terms of using the value of PM to make a decision on the need of adding more noise.

2.2.3 Differential Privacy

Following (Dwork, 2008), K is the randomized function applied to the actual data. D_1 and D_2 differ in at most one element; one is a proper subset of the other and the larger database contains just one additional row. A randomized function K gives ϵ-differential privacy if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq \text{Range}(K)$, the following applies:

$$\left|\log\left(\frac{Pr[K(D_1) \subseteq S]}{Pr[K(D_2) \subseteq S]}\right)\right| \leq \epsilon \tag{2.13}$$

This probability is taken over the distribution induced by K, and K satisfies the condition that the presence of absence of an individual's data will not affect their chance of receiving protection by more than a controlled amount as measured by ϵ. Selecting ϵ up to society and other stakeholders, with smaller values conferring greater protection (values like 0.01 are common). Note that $\epsilon = 0$ affords complete protection because the K function produces the same result for D_1 and D_2.

Note that the protection afforded by equation 2.13 is global in that it depends only on the Ds and not on who gets to see the data. However, if intruders have additional knowledge that is relevant to the information in the Ds, ϵ-level protection is not guaranteed. The formula can be extended to a group of data records being in or not in the dataset. For example, differing on at most c records produces $c\epsilon$-control rather than ϵ-control.

(Abowd and Vilhuber, 2008) address the situation wherein an intruder seeks to learn the value of Y_j for some record j in the dataset D. Let A be the prior information

that an intruder knows of dataset D, D^* is the released version of D, and let S represent information that an intruder knows about the generating process for D^*. Given (D^*, A, S), the intruder's density for Y_j is then

$$p(Y_j|D^*, A, S) \propto p(D^*|Y_j, A, S)p(Y_j|A, S) \tag{2.14}$$

This is known as a probabilistic risk measure. In Formula 2.14, $p(Y_j|A, S)$ is the intruder's prior distribution on Y_j based on (A, S) and D^* is used to sharpen the intruder's prior beliefs about Y_j. Under this formulation, records with high probabilities are at higher risk of disclosure. The requirement for this approach to be effective is that we should know what an intruder knows about the data A, which is a difficult assessment and risk increases if an intruder knows more than A. Additionally, it will be difficult to compute Formula 2.14 with a complex D^*,

Another approach is to release the data in tabular form. Table 2.2 is another example of swapping only one variable disease. Table 2.3 is its tabular version. Discussed by (Schlörer, 1981), tabulated data can be released to show the existence of swapping without actually identifying the respondents.

Original data set			
Number	Age	Male	Educate
1	24	1	1
2	26	0	3
3	20	1	3
4	25	1	3
5	23	0	2

Swapped data set		
Age	Male	Educate
20	1	1
24	0	3
23	1	3
25	1	3
26	0	2

Table 2.1: Swapping data for the Age variable for $N = 5$ individuals and $V = 3$ variables. Original data is on the left and the swapped data is on the right.

Original Data			
Number	Disease	Male	Race
1	0	1	1
2	1	0	1
3	1	1	0
4	0	1	0
5	1	0	1

Swapped data set		
Disease	Male	Race
1	1	1
0	0	1
0	1	0
0	1	0
1	0	1

Table 2.2: Swapping data for the Disease variable for $N = 5$ individuals and $V = 3$ variables. Original data is on the left and the swapped data is on the right.

Original data set							Swapped data set					
	Race							Race				
	0			1				0			1	
	Male			Male				Male			Male	
Disease	0	1	Disease	0	1		Disease	0	1	Disease	0	1
0	0	1	0	0	1		0	0	2	0	0	0
1	0	0	1	1	0		1	0	0	1	1	

Table 2.3: Tabular versions of original and swapped data from Table 2.2, in which it is based on two levels of varaibles including Disease, Male and Race. Original data is on the left and the swapped data is on the right.

n_{111}	n_{121}	n_{1+1}
n_{211}	n_{221}	n_{2+1}
n_{311}	n_{321}	n_{3+1}
n_{+11}	n_{+21}	n_{++1}

n_{111}	$n_{121}-1$	$n_{1+1}-1$
n_{211}	n_{221}	n_{2+1}
$n_{311}+1$	n_{321}	$n_{3+1}+1$
$n_{+11}+1$	$n_{+21}+1$	n_{++1}

n_{112}	n_{122}	n_{1+2}
n_{212}	n_{222}	n_{2+2}
n_{312}	n_{322}	n_{3+2}
n_{+12}	n_{+22}	n_{++2}

n_{112}	$n_{122}+1$	$n_{1+2}+1$
n_{212}	n_{222}	n_{2+2}
$n_{312}-1$	n_{322}	$n_{3+2}-1$
$n_{+12}-1$	$n_{+22}+1$	n_{++2}

Table 2.4: Data swapping in a Three-way contingency table with entries n_{ij}. The original table is on the left and the table with observations from the (1,2,1) amd (3,1,2) cells, swapped between layers on the right.

Chapter 3

Data Analysis and Validation

3.1 Background

As reported by (Elmer et al., 2006), high blood pressure (BP) is an important risk factor for cardiovascular disease (CVD). Current national recommendations for the prevention of high blood pressure emphasize, "lifestyle modification" with weight loss, physical activity, and alcohol consumption are all considered as lifestyle variables. Arterial blood pressure is sometimes referred as blood pressure (BP). The blood pressure in the circulation is a result of pumping of the heart. During each heartbeat, blood pressure varies between a maximum (systolic) and a minimum (diastolic) pressure. The Dietary Approaches to Stop Hypertension (DASH) diet can also help lower blood pressure. Participants in the trial are randomized to three intervention groups: "established," "established plus DASH," and "advice only." We want to assess different lifestyle factors that can help reduce blood pressure.

3.2 Application of a noise perturbation method

To get an idea of how well the noise addition technique works in practice, we test it with the actual data. To generate the noise multipliers, we experiment with several distributions listed below:

- In order to get a non-negative noise, we can either multiply a log-normal $(0, \sigma^2)$ noise or add a Normal (or Gaussian) noise. For normal (or Gaussian) noise, we denote the normal distribution as $N(1 \pm \mu, \sigma^2)$. Then we can assign μ with 0.1 and σ with 0.01, we would have two normal distribution with $N(0.9, 0.01^2)$ and $N(1.1, 0.01^2)$.
- Scaled beta distribution denoted as $A \times Beta(\alpha, \beta) + B$. Assign α with 2, β with 4, A with 0.1, and B with 1.1 or 0.9.
- Split triangle density function discussed in (Massell and Funk, 2007b): $f(x) = 0$ when $0.9 \leq x \leq 1.1, f(x) = (-k) \times (x - 1.2)$ when $1.1 \leq x \leq 1.2, f(x) = k \times (x - 0.8)$ when $0.8 \leq x \leq 0.9$, and $f(x) = 0$ otherwise.

After generating the noise multiplier, we apply the noise multiplier to the actual data using formula $z_i = y_i + r_i$ in Section 2.1.3. We standardize the data set to a common scale so that we can add the same magnitude of noise to each variable when we apply the noise perturbation method. After that, we examine multicollinearity, the situation in which two or more explanatory variables are highly correlated. Since

correlation between BMI and Weight is fairly strong with a value of 0.82, and correlation between BMI and Waist Circumference is strong with a value of 0.85, we drop Weight and Waist Circumference from our analysis. We also clean the data and convert some variables to factors rather than integers.

Figure 3.3 shows that the normal noise is symmetric and bell shaped. The beta distributions are slightly skewed to the right. The split triangle noise is distributed in two parts, one part ranging from 0.8 to 0.9 and the other part ranging from 1.1 to 1.2. All the noise distributes within the interval (0.8, 1.2).

Summary statistics of the actual data are listed in Table 3.1. Comparing them with those in Tables 3.2–3.6, we see that the summary statistics of modified data are generally higher than those of the actual data. This is as expected since the noise multipliers we used are all positive. Adding noise with mean larger than 0 will help to decrease the disclosure risk but will essentially introduce more bias than adding noise with mean 0. For Tables 3.2–3.5, the standard deviations are very close to the corresponding standard deviations of the actual data. For Table 3.6, the standard deviations are slightly higher than those of the actual data. Figure 3.1–3.2 illustrates the barplot to help easily compare results across tables.

A gross comparison of the space covered by the predictors from the actual data set and the new set can be made using routine dimension reduction techniques such as principle component analysis. If the actual data and the modified data are generated from the same mechanism, then the projection of these data will overlap in the scatter

plot. However, if the actual data and modified data occupy different parts of the scatter plot, then they may not be generated by the same mechanism and predictions for the modified data and actual data might be different. Therefore, we can use this way to examine the similarity of the actual data and modified data.

Figures 3.4–3.8 display the scatter plot of projection of the actual data and modified data onto the first two principal components. In this case, the actual data and modified data appear to occupy the same space as determined by these components for all five types of noise.

Next, we explore the similarity between actual data and modified data from another aspect. We run the regression model using the actual data and perturbed data and compare results. Since our data are gathered over time on the same individuals for follow up visits at 0, 3, 6, 12, 18 months, we have multiple measures per subject. Multiple measures from one subject are dependent, therefore we include a random subject effect, which induces longitudinal correlation by allowing a different "baseline" SBP or DBP for each subject. Since SBP and DBP are two different measures, we model them separately. We use the `R` statistical programming language `http://www.r-project.org/` for the analysis; `nlme` (Pinheiro et al., 2014) is used to fit the linear mixed effects models. The response is the diastolic blood pressure change from baseline and the systolic blood pressure change from baseline respectively for model 1 and model 2. The predictors are Baseline SBP, Baseline DBP, BMI, AGE, SEX, Treatment, and Follow-up visits. Since the distribution of our data

are fairly skewed, we generated 500 bootstrap samples out of the original data. For a given iteration of bootstrap re-sampling, the model is built on the selected samples to estimate the random effect. We then got the percentile interval using the 2.5% percentile and 97.5% percentile of the empirical distribution of bootstrap estimates for standard error.

The mixed model results are shown in Table 3.7. Model 1 the for SBP and Model 2 for DBP. For both models, SBP_Base (SBP at baseline), DBP_Base (DBP at baseline), FV (follow up visit), and TX (treatment) are highly significant. Sex is a significant term for Model 1 but not significant for Model 2.

Other than the split triangle noise, AICs from the rest four noise perturbed data are very similar from the actual data. Split triangle noise tends to have a larger AIC value. The estimates of slope from all the data sets are close (see Tables 3.7–3.12). The estimates of regression intercept are larger than those of the actual data which are to expect since all our noise are positive. Besides, the split triangle noise gives us a slightly different estimated slope.

As discussed previously in Section 2.2.2, it's desirable to measure the amount of under-perturbation of sensitive cells and over perturbation of safe cells. We use Formulas 2.10–2.11 to calculate the alpha measures and beta measures. We assume all cells are sensitive and set prot= 0 for easy calculation. Table 3.13 displays the value of alpha measures and beta measures across the five types of noise. For alpha measures, we can see that noise from $0.1 * Beta(2, 4) + 1.1$ distribution have the

least value and noise from $N(0.9, 0.01^2)$ distribution have the largest value. For beta measures, split triangle distribution have the least value and $0.1 * Beta(2, 4) + 1.1$ have the largest value. We cam see that modified data with smaller alpha measure values are associated with a relatively larger beta measure values and vice versa. The smallest values we got when we sum up the alpha measure and beta measure values are distribution $N(1.1, 0.01^2)$ and $0.1 * Beta(2, 4) + 1.1$. If we want to minimize both alpha measure and beta measures, noise generated from those two noise distribution might be better in general.

3.3 Differential privacy

As discussed in Section 2.2.3, we examine the differential privacy feature by calculating ϵ based on Formula 2.13 using the same data. We denote the median of variable age as "m_{orig};" and the standard deviation as σ_α in the actual data. The noise multipliers added to age in the actual data are normally distributed with mean 0, and standard deviation $\sigma_\delta, \sigma_\beta, \sigma_\gamma$, where $\sigma_\delta = \sigma_\alpha$, $\sigma_\beta = \sigma_\alpha/2$, and $\sigma_\gamma = 2\sigma_\alpha$. We characterize the three types of noise as low, moderate, and high noise respectively. We generated noise and added the noise to the actual data. We then repeat this process for 200 times.

We compute the following two separate evaluations:

- The number of times that the median is equal to m_{orig}
- The number of times that the mean is close to m_{orig}, closeness are defined as

means $\pm$ 0.2

Figure 1 shows the histogram of three types of noise. Low noise are ranging from -15 to 15, moderate noise are ranging from -30 to 30, and high noise are ranging from -55 to 55. The variance is the biggest for high noise and the lowest for low noise. High level of noise will be able to get us more protection, which means lower disclosure risk and vice versa.

According to the foregoing definition, D_1 is the original data and D_2 is the modified data, and several approaches were taken to modify the data and see how it affects the ϵ value in Formula 2.13. For example, the maximum data value can be removed from the original data, and, separately, the minimum and calculate the probability denoted by z, take logs and get ϵ value.

Tables 3.14-3.15 show the ϵ value when we add three types of noise by removing the max, minimum, and the median according to two different evaluations.

3.4 Synthetic data method

Here we explore the synthetic data method discussed in Section 2.1.4 and show how it works. The first task is to fit a multiple regression model and the model we fit is $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \epsilon_i$. We assume that those ϵ_i's are statistically independent, each with mean 0 and standard deviation σ. In our data set, the response is systolic blood pressure (SBP), X_1 being the SBP at baseline

(SBP_BASE), X_2 being the Age variable, and X_3 being the BMI variable. Table 3.16 is the summary statistics of fitted model and Table 3.17 is the anova table of fitted model.

Second, we randomly draw n observations from variable AGE, BMI, and SBP_BASE independently with replacement, where n equal to the total observations in the data set. Table 3.18 shows the five summary statistics of original and synthetic for Age, BMI, and SBP_BASE variables. Figures 3.10–3.12 show the histogram of original and synthetic variables (Age, BMI, and SBP_BASE).

Lastly, we use the synthetic variables Age, BMI, SBP_BASE and the estimated regression slope and intercept from the fitted model to predict the values of systolic blood pressure(SBP). The summary statistics are shown in Table 3.21. Though the minimum and maximum of synthetic variables SBP's are not close to the actual values, the mean and median are fairly close to the actual variables values. We can improve the performance of synthetic data by fitting a more saturated model. We can also use other sampling method like bootstrapping to sample the variables.

Since we independently sample covariate from the actual data set, we might distort the correlation between covariates. Another way is to sample covariates all together and we'll therefore preserve the original correlation between covariates. We'll apply the same procedures and see how it differs. See Tables 3.20–3.21 for the summary statistics.

We can see that two sampling methods give us the similar result. That might be

because the correlation between the explanatory variables is small. If we get a larger correlation between the explanatory variables, we'll expect to see a larger difference.

Variables	Min	Max	Mean	St. Dev.	Median
Baseline SBP	115.8	160.5	134.9	9.6	134.2
Baseline DBP	73.8	98.0	84.8	4.2	84.4
DBP	41.0	122.0	80.8	7.8	81.0
SBP	85.0	183.0	128.5	12.7	128.0
BMI	21.0	48.2	33.1	5.8	32.3
AGE	25.0	79.0	50.0	8.9	50.0

Table 3.1: Summary statistics of the actual (micro) data for $V = 4$ variables.

Variables	Min	Max	Mean	St. Dev.	Median
Baseline SBP	126.0	171.0	145.4	9.6	145.0
Baseline DBP	78.3	102.7	89.4	4.2	89.0
DBP	49.5	130.5	89.3	7.8	89.6
SBP	99.0	196.9	142.4	12.7	141.8
BMI	27.3	54.6	39.4	5.8	38.6
AGE	34.7	88.9	59.7	8.9	59.7

Table 3.2: Summary statistics of the modified data. The noise comes from the distribution $N(0.9, 0.01^2)$ and the noise was then added to the actual (micro) data.

Variables	Min	Max	Mean	St. Dev.	Median
Baseline SBP	124.0	169.0	143.5	9.6	143.0
Baseline DBP	77.5	101.8	88.5	4.2	88.1
DBP	48.0	129.0	87.7	7.8	88.0
SBP	96.4	194.5	139.9	12.7	139.2
BMI	26.2	53.4	38.3	5.8	37.5
AGE	32.9	87.1	58.0	8.9	57.9

Table 3.3: Summary statistics of the modified data. The noise comes from the distribution $N(1.1, 0.01^2)$ and the noise was then added to the actual (micro) data.

Variables	Min	Max	Mean	St. Dev.	Median
Baseline SBP	124.0	170.0	143.8	9.6	143.0
Baseline DBP	77.5	101.9	88.7	4.2	88.3
DBP	48.6	129.3	88.0	7.8	88.3
SBP	96.8	194.7	140.3	12.7	139.6
BMI	26.4	53.7	38.5	5.8	37.7
AGE	33.1	87.5	58.3	8.9	58.1

Table 3.4: Summary statistics of the modified data. The noise comes from the distribution $0.1 * Beta(2, 4) + 0.9$ and the noise was then added to the actual (micro) data.

Variables	Min	Max	Mean	St. Dev.	Median
Baseline SBP	126.0	172.0	145.7	9.6	145.0
Baseline DBP	78.3	102.7	89.5	4.2	89.1
DBP	50.1	130.9	89.6	7.8	89.8
SBP	99.3	197.2	142.9	12.7	142.1
BMI	27.5	54.8	39.6	5.8	38.8
AGE	34.9	89.3	60.0	8.9	59.9

Table 3.5: Summary statistics of the modified data. The noise comes from the distribution $0.1 * Beta(2, 4) + 1.1$ and the noise was then added to the actual (micro) data.

Variables	Min	Max	Mean	St. Dev.	Median
Baseline SBP	124.0	171.0	144.5	9.7	144.0
Baseline DBP	77.2	102.9	88.9	4.2	88.6
DBP	49.7	128.9	88.5	7.9	88.9
SBP	100.0	197.3	141.2	12.8	140.5
BMI	26.1	55.0	38.9	5.9	38.2
AGE	32.4	89.5	58.8	9.0	58.4

Table 3.6: Summary statistics of the modified data. The noise comes from the split triangle ditribution and the noise was then added to the actual (micro) data.

	Model 1 (SE)	Model 2 (SE)
(Intercept)	0.57 (0.05)	0.61 (0.05)
Baseline SBP	0.60 (0.03)	
BMI	0.01 (0.02)	−0.01 (0.02)
AGE	−0.03 (0.02)	−0.04 (0.02)
Hypertension	−0.01 (0.05)	−0.02 (0.05)
SEX	0.08 (0.04)	0.03 (0.04)
Treatment2	−0.15 (0.04)	−0.13 (0.05)
Treatment3	−0.18 (0.04)	−0.19 (0.05)
Follow-up visit3	−0.54 (0.03)	−0.55 (0.04)
Follow-up visit6	−0.74 (0.03)	−0.67 (0.04)
Follow-up visit12	−0.56 (0.03)	−0.62 (0.04)
Follow-up visit18	−0.67 (0.03)	−0.75 (0.04)
Baseline DBP		0.45 (0.02)
AIC	8908.88	9814.75
Random effects (95% Confidence Interval)		
Residual	$(0.57, 0.61)$	$(0.63, 0.68)$

Table 3.7: Estimates of fixed effects using actual data. Model 1 including all the variables except Baseline DBP and Model 2 including all the variables except Baseline SBP.

	Model 1 (SE)	Model 2 (SE)
(Intercept)	0.94 (0.05)	1.14 (0.06)
Baseline SBP	0.60 (0.03)	
BMI	0.01 (0.02)	−0.01 (0.02)
AGE	−0.03 (0.02)	−0.04 (0.02)
Hypertension	−0.01 (0.05)	−0.02 (0.05)
SEX	0.08 (0.04)	0.03 (0.04)
Treatment2	−0.15 (0.04)	−0.13 (0.05)
Treatment3	−0.18 (0.04)	−0.19 (0.05)
Follow-up visit3	−0.54 (0.03)	−0.55 (0.04)
Follow-up visit6	−0.74 (0.03)	−0.67 (0.04)
Follow-up visit12	−0.56 (0.03)	−0.62 (0.04)
Follow-up visit18	−0.67 (0.03)	−0.75 (0.04)
Baseline DBP		0.45 (0.02)
AIC	8909.71	9815.13
Random effects (95% Confidence Interval)		
Residual	$(0.57, 0.60)$	$(0.63, 0.68)$

Table 3.8: Estimates of fixed effects using modified data. The noise distribution comes from $N(0.9, 0.01^2)$ and the noise was then added to the actual (micro) data. Model 1 including all the variables except Baseline DBP and Model 2 including all the variables except Baseline SBP.

	Model 1 (SE)	Model 2 (SE)
(Intercept)	1.03 (0.05)	1.26 (0.06)
Baseline SBP	0.60 (0.03)	
BMI	0.01 (0.02)	−0.01 (0.02)
AGE	−0.03 (0.02)	−0.04 (0.02)
Hypertension	−0.01 (0.05)	−0.02 (0.05)
SEX	0.08 (0.04)	0.03 (0.04)
Treatment2	−0.15 (0.04)	−0.13 (0.05)
Treatment3	−0.19 (0.04)	−0.19 (0.05)
Follow-up visit3	−0.54 (0.03)	−0.55 (0.04)
Follow-up visit6	−0.74 (0.03)	−0.67 (0.04)
Follow-up visit12	−0.56 (0.03)	−0.62 (0.04)
Follow-up visit18	−0.67 (0.03)	−0.75 (0.04)
Baseline DBP		0.45 (0.02)
AIC	8906.10	9817.04
Random effects (95% Confidence Interval)		
Residual	$(0.57, 0.61)$	$(0.63, 0.68)$

Table 3.9: Estimates of fixed effects using modified data. The noise distribution comes from $N(1.1, 0.01^2)$ and the noise was then added to the actual (micro) data. Model 1 including all the variables except Baseline DBP and Model 2 including all the variables except Baseline SBP.

	Model 1 (SE)	Model 2 (SE)
(Intercept)	0.96 (0.05)	1.16 (0.06)
Baseline SBP	0.60 (0.03)	
BMI	0.01 (0.02)	0.00 (0.02)
AGE	−0.03 (0.02)	−0.03 (0.02)
Hypertension	−0.01 (0.05)	−0.02 (0.05)
SEX	0.08 (0.04)	0.03 (0.04)
Treatment2	−0.15 (0.04)	−0.13 (0.05)
Treatment3	−0.18 (0.04)	−0.19 (0.05)
Follow-up visit3	−0.54 (0.03)	−0.55 (0.04)
Follow-up visit6	−0.74 (0.03)	−0.67 (0.04)
Follow-up visit12	−0.56 (0.03)	−0.62 (0.04)
Follow-up visit18	−0.67 (0.03)	−0.75 (0.04)
Baseline DBP		0.45 (0.02)
AIC	8908.80	9816.75
Random effects (95% Confidence Interval)		
Residual	$(0.57, 0.60)$	$(0.63, 0.68)$

Table 3.10: Estimates of fixed effects using modified data. The noise distribution comes from $0.1 * Beta(2, 4) + 0.9$ and the noise was then added to the actual (micro) data. Model 1 including all the variables except Baseline DBP and Model 2 including all the variables except Baseline SBP.

	Model 1 (SE)	Model 2 (SE)
(Intercept)	1.04 (0.05)	1.28 (0.06)
Baseline SBP	0.60 (0.03)	
BMI	0.01 (0.02)	0.00 (0.02)
AGE	−0.03 (0.02)	−0.03 (0.02)
Hypertension	−0.01 (0.05)	−0.02 (0.05)
SEX	0.08 (0.04)	0.03 (0.04)
Treatment2	−0.15 (0.04)	−0.13 (0.05)
Treatment3	−0.18 (0.04)	−0.19 (0.05)
Follow-up visit3	−0.54 (0.03)	−0.55 (0.04)
Follow-up visit6	−0.74 (0.03)	−0.67 (0.04)
Follow-up visit12	−0.56 (0.03)	−0.62 (0.04)
Follow-up visit18	−0.67 (0.03)	−0.75 (0.04)
Baseline DBP		0.45 (0.02)
AIC	8908.80	9816.75
Random effects (95% Confidence Interval)		
Residual	$(0.57, 0.61)$	$(0.63, 0.68)$

Table 3.11: Estimates of fixed effects using modified data. The noise distribution comes from $0.1 * Beta(2, 4) + 1.1$ and the noise was then added to the actual (micro) data. Model 1 including all the variables except Baseline DBP and Model 2 including all the variables except Baseline SBP.

	Model 1 (SE)	Model 2 (SE)
(Intercept)	0.99 (0.05)	1.23 (0.06)
Baseline SBP	0.54 (0.02)	
BMI	0.01 (0.02)	0.00 (0.02)
AGE	−0.01 (0.02)	−0.04 (0.02)
Hypertension	0.06 (0.05)	0.02 (0.05)
SEX	0.08 (0.04)	0.01 (0.04)
Treatment2	−0.15 (0.04)	−0.13 (0.05)
Treatment3	−0.17 (0.04)	−0.19 (0.05)
Follow-up visit3	−0.54 (0.03)	−0.54 (0.04)
Follow-up visit6	−0.75 (0.03)	−0.67 (0.04)
Follow-up visit12	−0.56 (0.03)	−0.61 (0.04)
Follow-up visit18	−0.67 (0.03)	−0.74 (0.04)
Baseline DBP		0.42 (0.02)
AIC	9114.87	9977.69
Random effects (95% Confidence Interval)		
Residual	$(0.59, 0.63)$	$(0.64, 0.69)$

Table 3.12: Estimates of fixed effects using modified data. The noise comes from split triangle distribution and the noise was then added to the actual (micro) data. Model 1 including all the variables except Baseline DBP and Model 2 including all the variables except Baseline SBP.

	Alpha measure	Beta measure
$N(0.9, 0.01^2)$	0.25	0.65
$N(1.1, 0.01^2)$	0.08	0.78
$0.1 * Beta(2, 4) + 0.9$	0.22	0.67
$0.1 * Beta(2, 4) + 1.1$	0.06	0.80
Split triangle distribution	0.18	0.71

Table 3.13: Alpha and Beta measures for the five modified data. The left column represents the noise distribution and the noise was then added to the actual (micro) data resulting the five moidified data.

	Low Noise	Moderate Noise	Strong Noise
Removing the max	0.005	0.006	0.084
Removing the minimum	0.005	0.024	0.049
Removing the median	0.020	0.006	0.030

Table 3.14: Examine three types of noise, i.e. low noise, moderate noise and strong noise. Using the evaluation criteria that median is the same as median m_{orig}), ϵ values were then calculated by removing the max, minimum and median respectively.

	Low Noise	Moderate Noise	Strong Noise
Removing the max	0.005	0.110	0.033
Removing the minimum	0.049	0.010	0.097
Removing the median	0.026	0.054	0.097

Table 3.15: Examine three types of noise, i.e. low noise, moderate noise and strong noise. Using the evaluation criteria that means is close to $m_{orig} \pm$ 0.2, ϵ values were then calculated by removing the max, minimum and median respectively.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	0	0.0127	0.00	1.0000
Baseline SBP	0.5978	0.0134	44.74	0
AGE	-0.0308	0.0139	-2.21	0.0272
BMI	0.0123	0.0133	0.92	0.3553

Table 3.16: Regression coefficients using the actual (micro) data. The response is variable SBP and explanatory variables are Baseline SBP, AGE, BMI.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Baseline SBP	1	1405.23	1405.23	2154.66	0.0000
AGE	1	4.48	4.48	6.87	0.0088
BMI	1	0.56	0.56	0.85	0.3553
Residuals	4046	2638.73	0.65		

Table 3.17: ANOVA table for the actual (micro) data from the model 3.16. The response is variable SBP and explanatory variables are Baseline SBP, AGE, BMI.

	Age		BMI		Baseline SBP	
	Original	Synthetic	Original	Synthetic	Original	Synthetic
Min	-2.810	-2.810	-2.080	-2.080	-2.000	-2.000
1^{st} quantile	-0.670	-0.670	-0.754	-0.790	-0.800	-0.824
Median	0.004	0.004	-0.113	-0.137	-0.067	-0.119
Mean	0	0.009	0.264	0	0	-0.024
3^{rd} quantile	0.679	0.651	0.754	0.725	0.691	0.665
Max	3.264	3.264	2.602	2.602	2.680	2.680

Table 3.18: Summary statistics of original and synthetic data for variables Age, BMI and Baseline SBP respectively. Explanatory variables from synthetic data obtained by sample four variables separately.

	Min	1^{st} quantile	Median	Mean	3^{rd} quantile	Max
Original SBP	-3.44	-0.67	-0.04	0	0.62	4.31
Synthetic SBP	-1.26	-0.50	-0.04	-0.01	0.41	1.62

Table 3.19: Summary statistics of original and synthetic data for response SBP variables. Explanatory variables from synthetic data obtained by sampling four variables separately and response from synthetic data obtained by plugging into the actual regression coefficients estimates.

	Age		BMI		Baseline SBP	
	Original	Synthetic	Original	Synthetic	Original	Synthetic
Min	-2.810	-2.810	-2.080	-2.080	-2.000	-2.000
1^{st} quantile	-0.670	-0.670	-0.754	-0.792	-0.800	-0.824
Median	0.004	0.004	-0.113	-0.133	-0.067	-0.067
Mean	0	0.015	0.264	0.002	0	-0.007
3^{rd} quantile	0.679	0.679	0.754	0.708	0.691	0.691
Max	3.264	3.264	2.602	2.602	2.680	2.680

Table 3.20: Summary statistics of original and synthetic data for variables Age, BMI and Baseline SBP respectively. Explanatory variables from synthetic data obtained by sample four variables together.

	Min	1^{st} quantile	Median	Mean	3^{rd} quantile	Max
Original SBP	-3.44	-0.67	-0.04	0	0.62	4.31
Synthetic SBP	-1.15	-0.50	-0.04	-0.004	0.42	1.61

Table 3.21: Summary statistics of original and synthetic data for response SBP variables. Explanatory variables from synthetic data obtained by sampling four variables together and response from synthetic data obtained by plugging into the actual regression coefficients estimates.

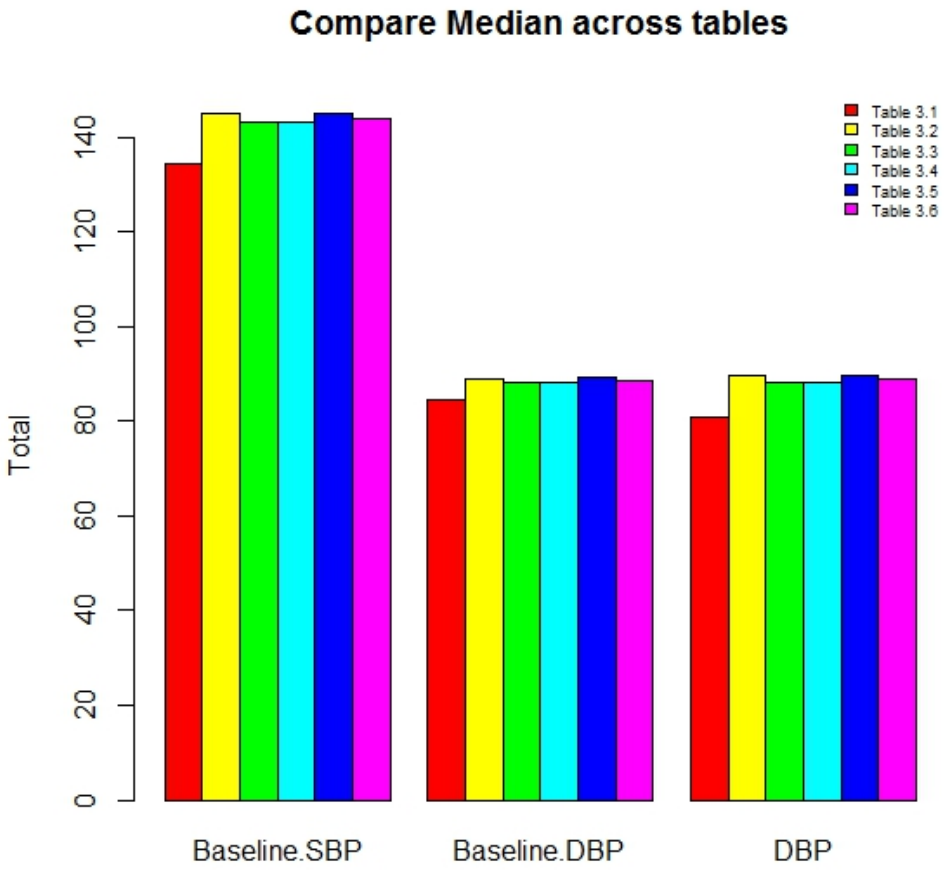

Figure 3.1: Compare median values across table 3.1–3.6 for variables Baseline SBP, Baseline DBP and DBP

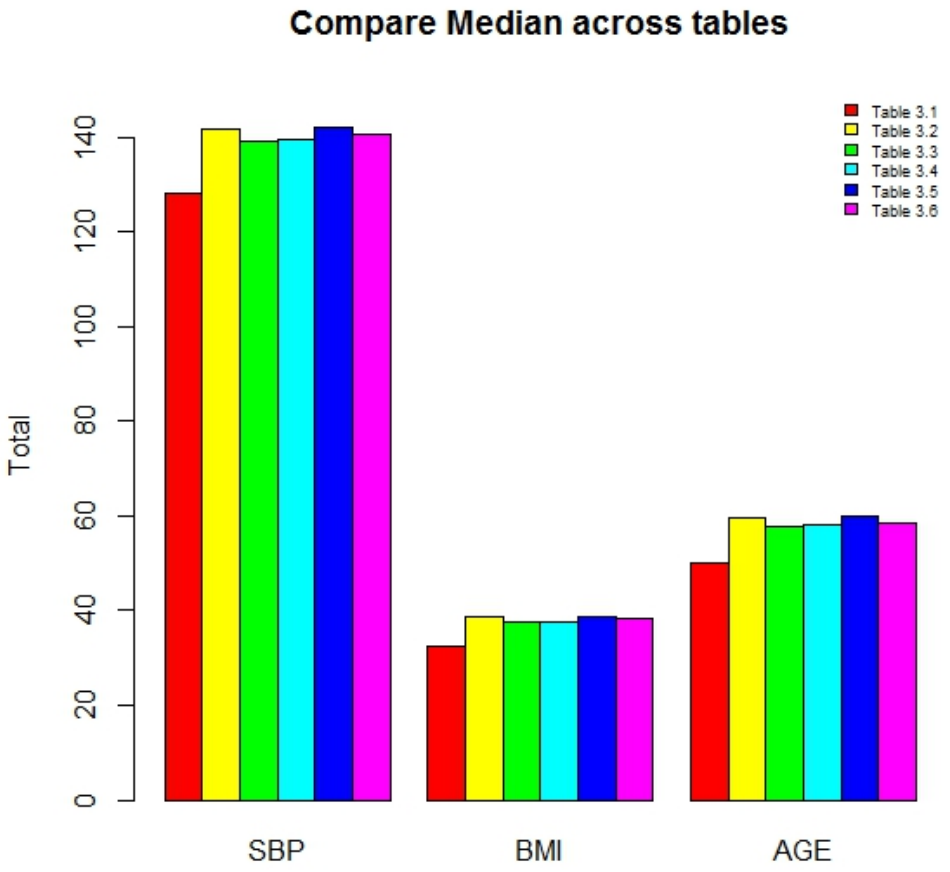

Figure 3.2: Compare median values across table 3.1–3.6 for variables SBP, BMI, and AGE

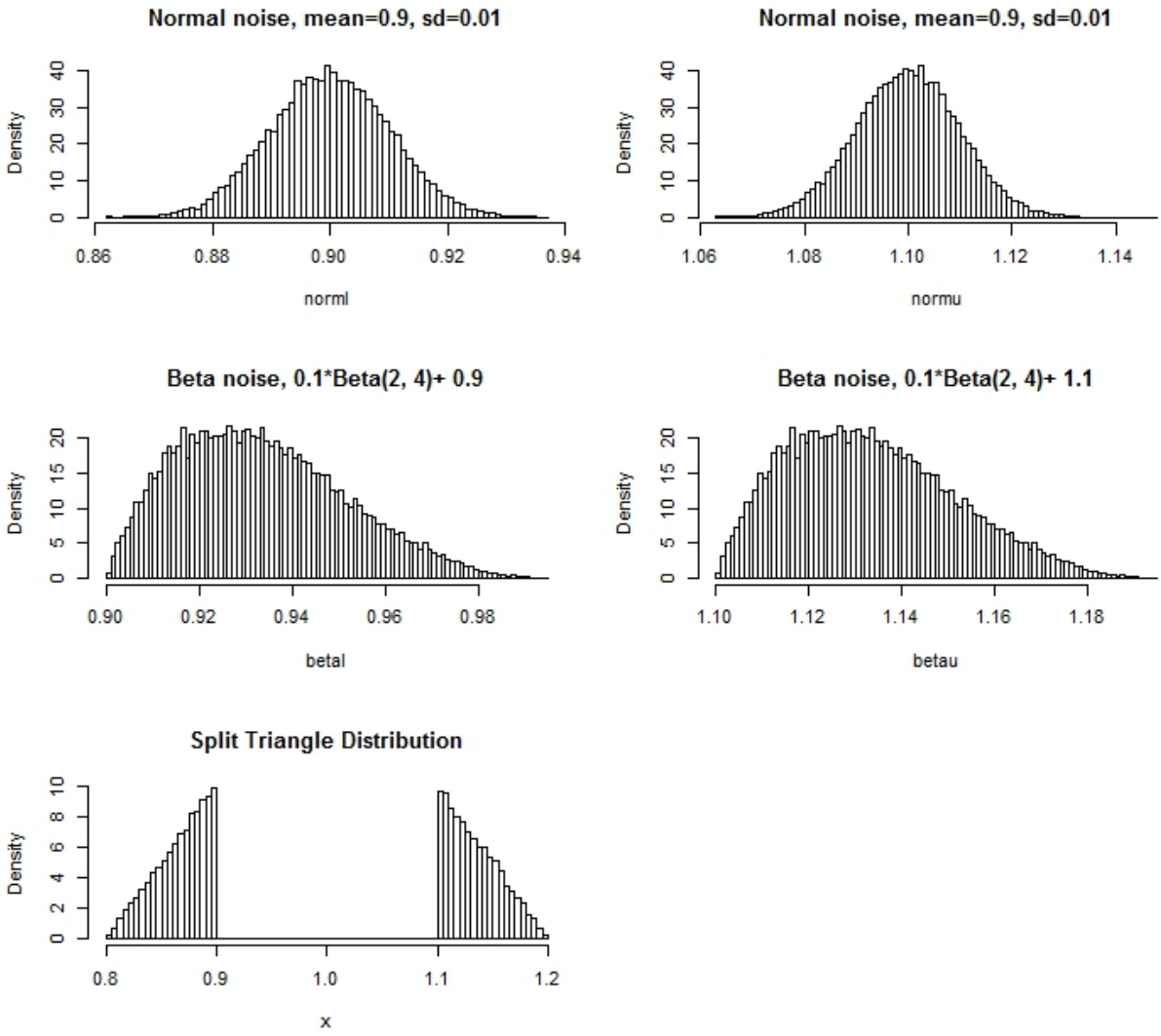

Figure 3.3: Histogram of 5 types of noise distribution. Modified data were then generated using those noise distributions.

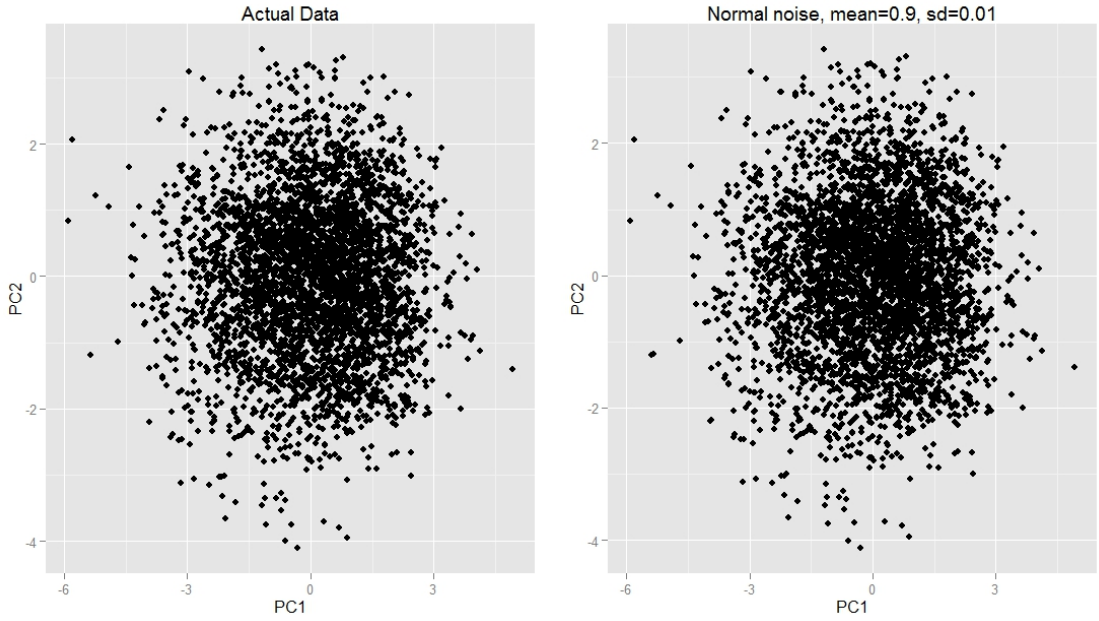

Figure 3.4: PCA plots for the actual data and modified data. The modified data were generated using noise coming from distribution $N(0.9, 0.01^2)$. The actual data are on the left and the modified data are on the right.

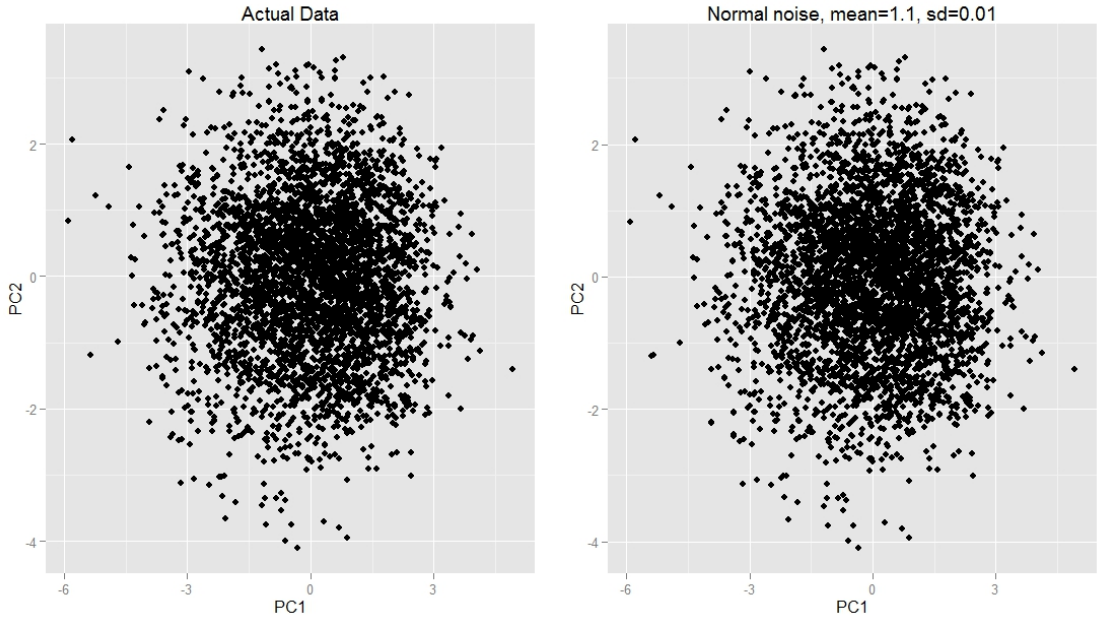

Figure 3.5: PCA plots for the actual data and modified data. The modified data were generated using noise coming from distribution $N(1.1, 0.01^2)$. The actual data are on the left and the modified data are on the right.

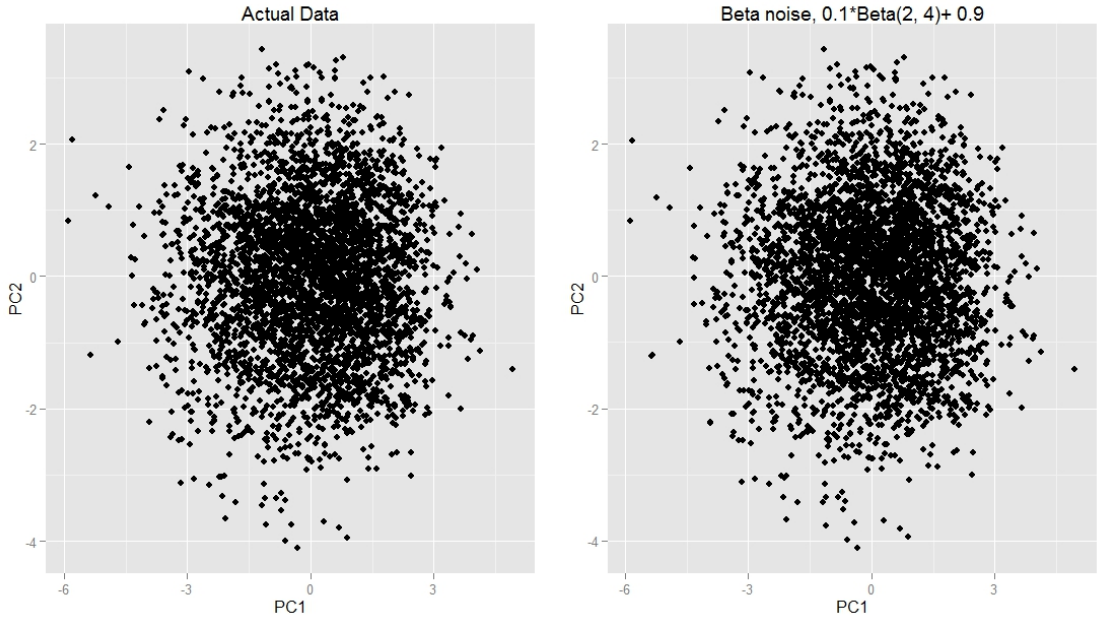

Figure 3.6: PCA plots for the actual data and modified data. The modified data were generated using noise coming from distribution $0.1 * Beta(2, 4) + 0.9$. The actual data are on the left and the modified data are on the right.

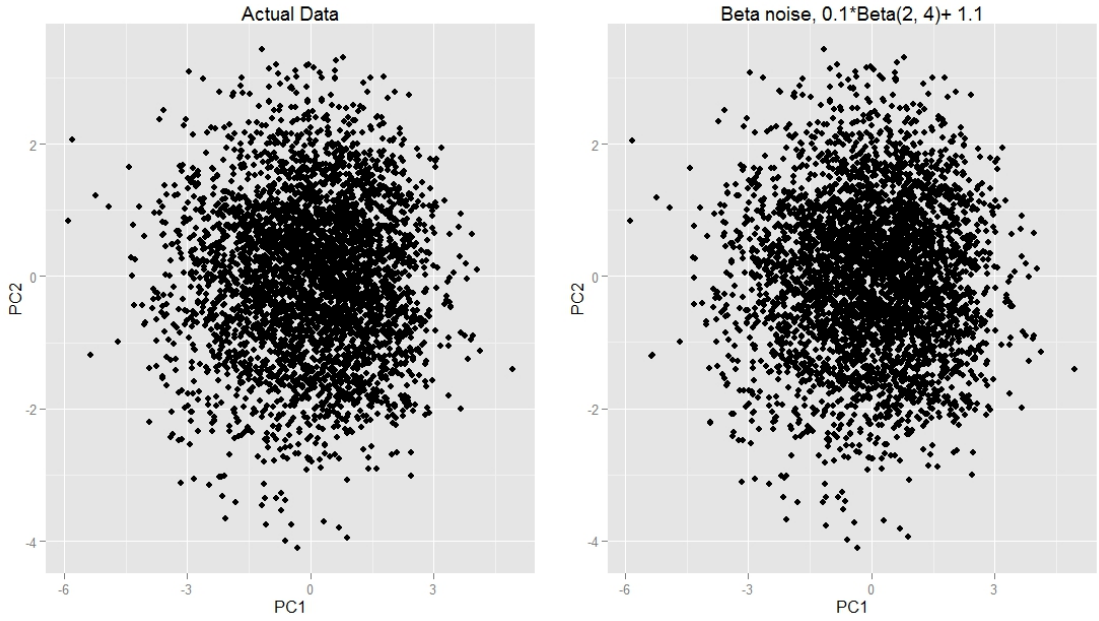

Figure 3.7: PCA plots for the actual data and modified data. The modified data were generated using noise coming from distribution $0.1 * Beta(2, 4) + 1.1$. The actual data are on the left and the modified data are on the right.

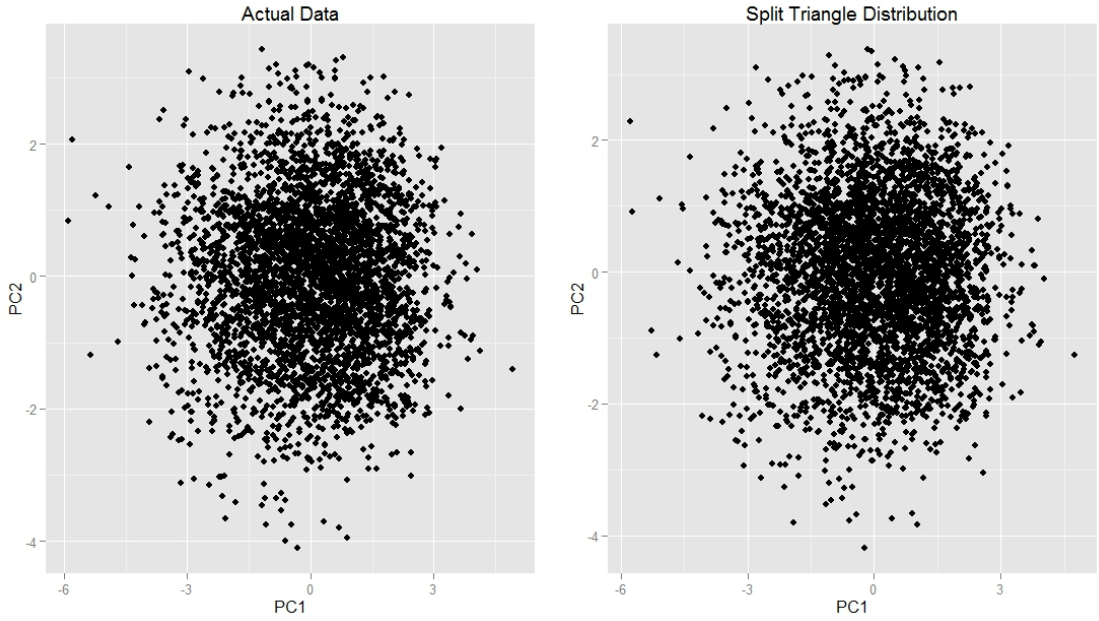

Figure 3.8: PCA plots for the actual data and modified data. The modified data were generated using noise coming from split triangle distribution. The actual data are on the left and the modified data are on the right.

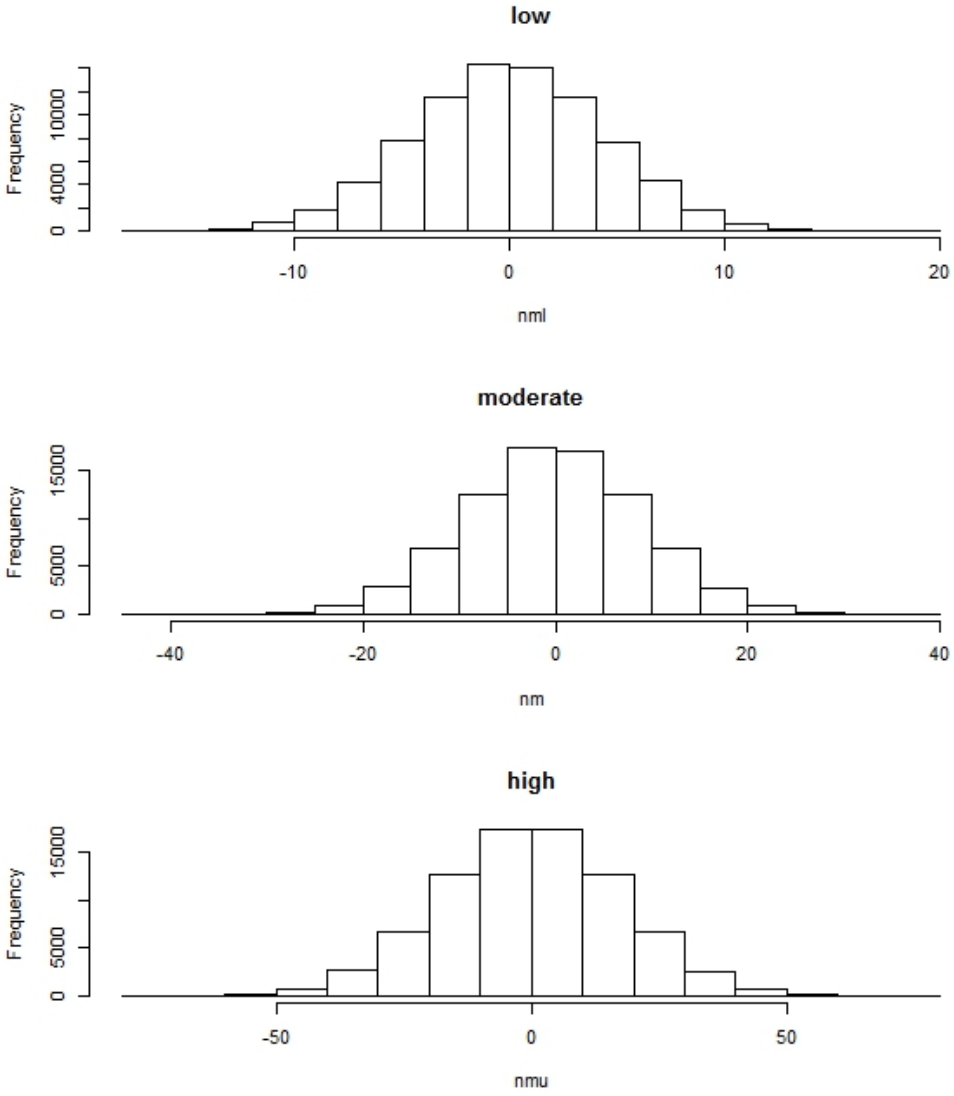

Figure 3.9: Histogram of three types of noise: low, moderate and high. The low, moderate, and high noise come from normal distribution.

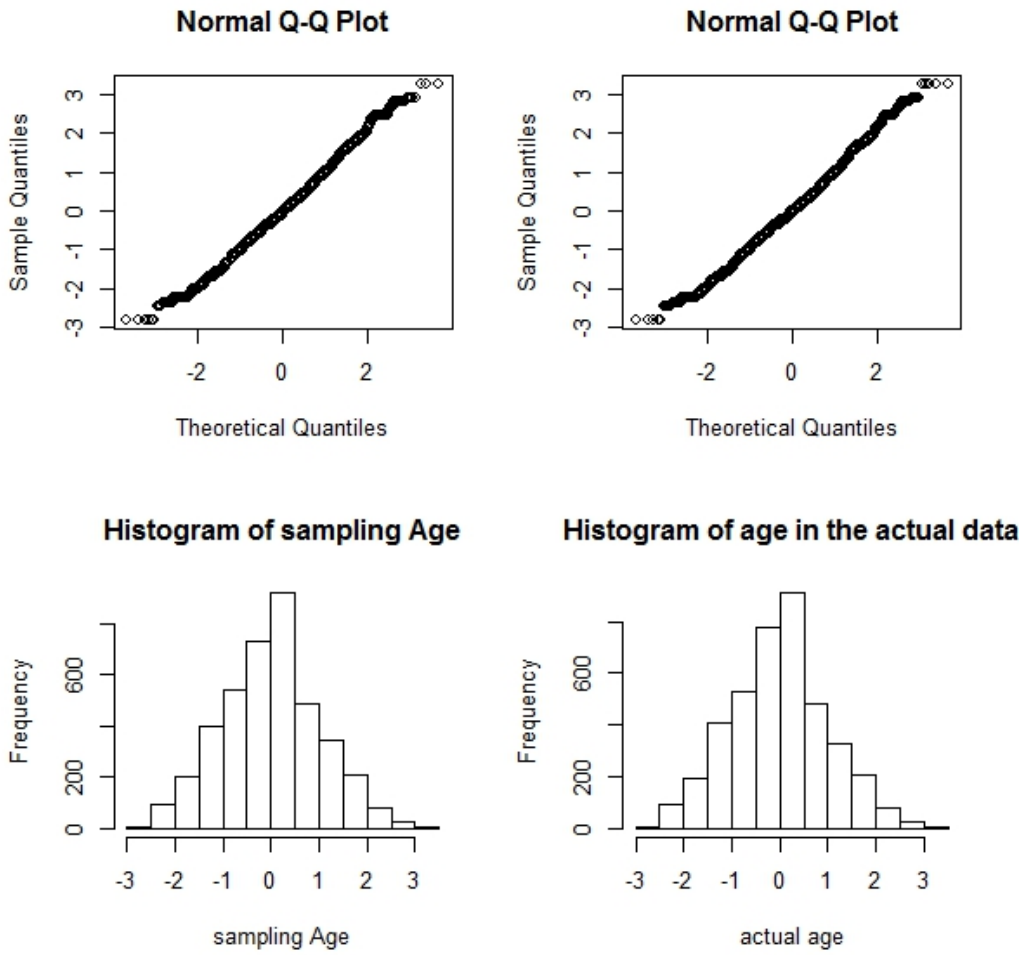

Figure 3.10: Histogram of variable AGE from actual and Sampled data. Sampled data are on the left and actual data are on the right.

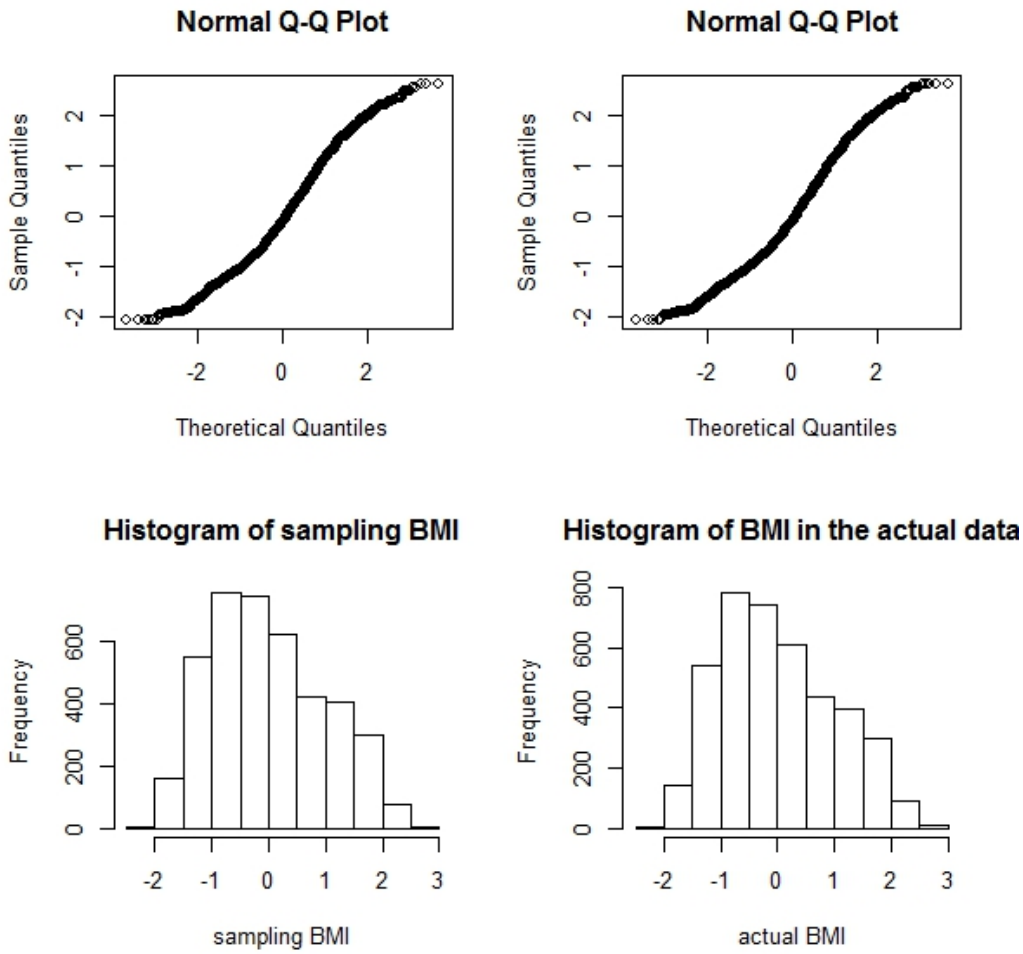

Figure 3.11: Histogram of variable BMI sampled from actual and Sampled data. Sampled data are on the left and actual data are on the right.

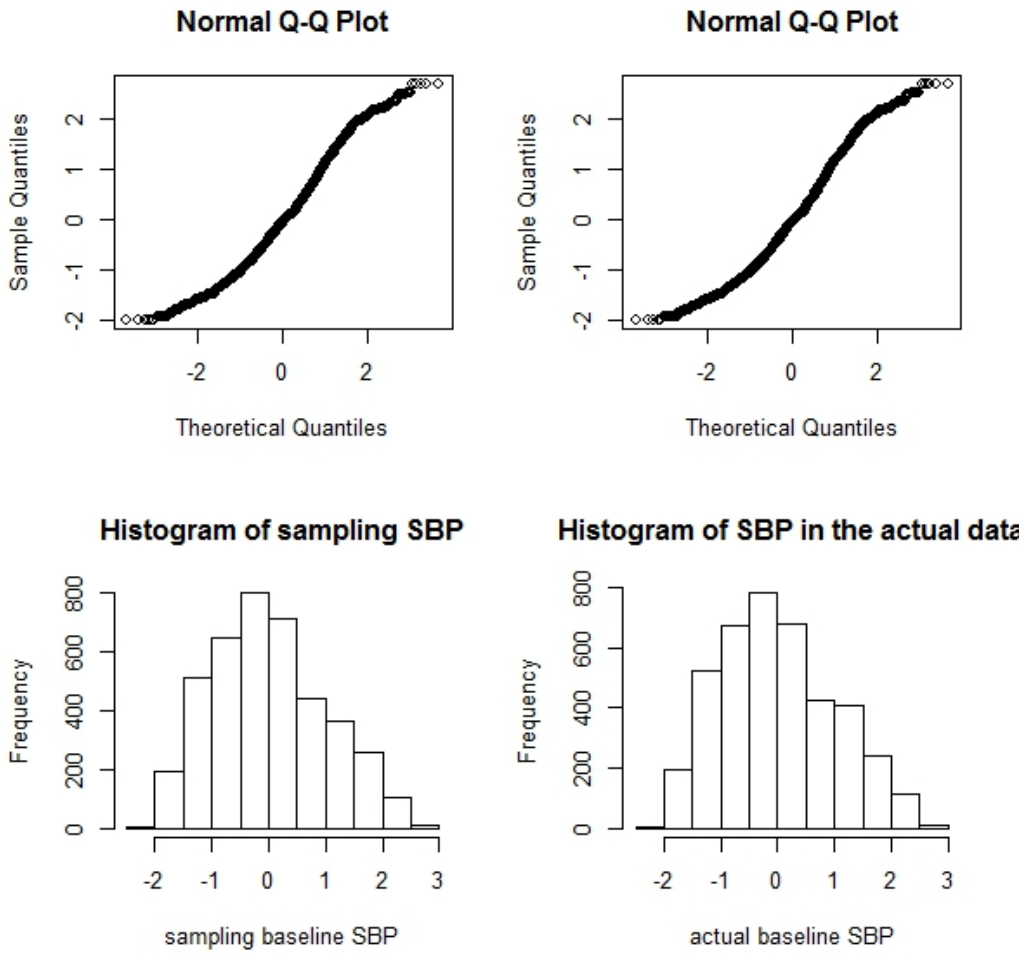

Figure 3.12: Histogram of variable Baseline SBP from actual and Sampled data. Sampled data are on the left and actual data are on the right.

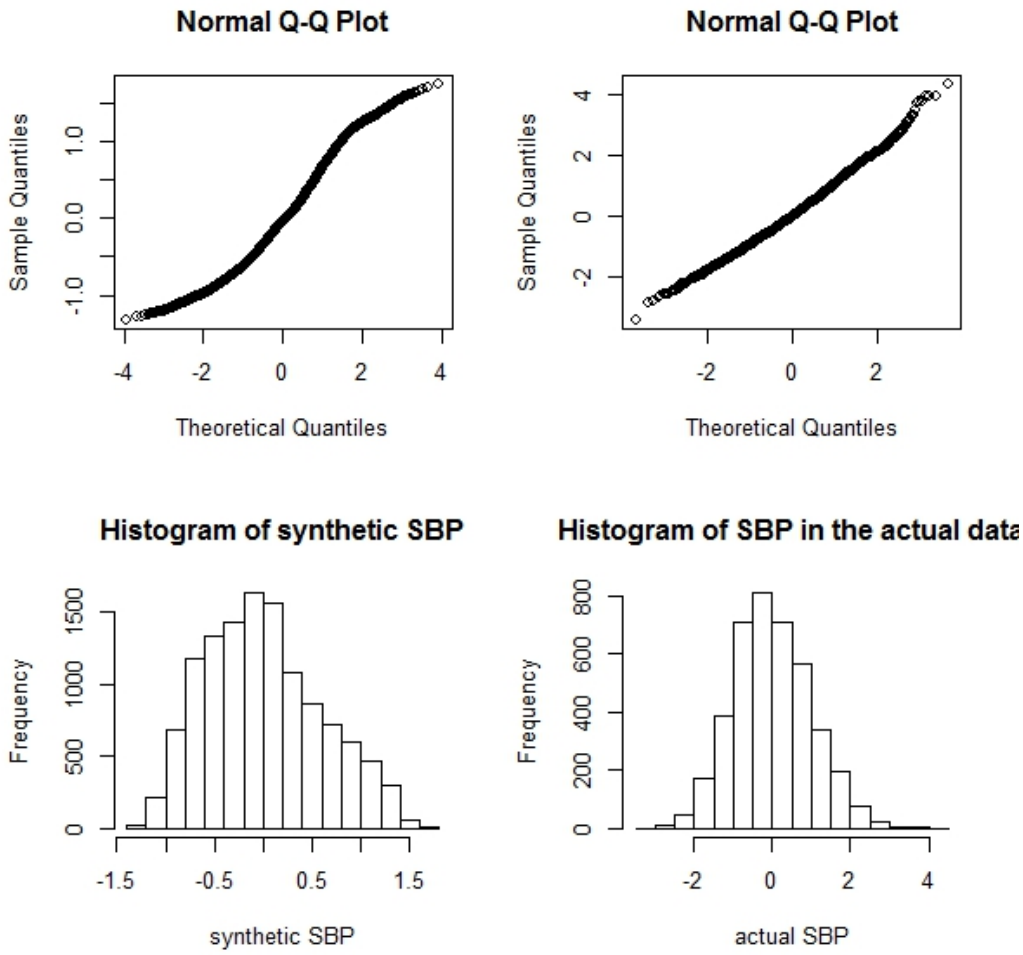

Figure 3.13: Histogram of response variable SBP from actual and Sampled data. Explanatory variables are sampled from actual data together and response variable SBP are obtained by plugging into the actual regression coefficients estimates. Sampled data are on the left and actual data are on the right.

Chapter 4

Discussion

In this study we have investigated performance of the noise perturbation and synthetic data methods for preserving confidentiality of private information in micro data. We generate different perturbations from several distributions and compare the differences among them. As shown in the foregoing figures, evidence from principal component analysis shows that modified data and actual data are generated from the same mechanism regardless of which noise distribution is used. Since the synthetic data come from the right regression model, we're expect to see very little difference for the regression coefficients between the actual data and synthetic data. For the noise perturbation method, all noise distributions produce very similar results and the results are fairly close to the actual (micro) data.

To assess the disclosure risk, we explore two different methods with micro generated data set. The advantage of the global scalar measures is that we got a value

for both alpha measure and beta measure for each generated data set. When we use this method, it would be important to decide a acceptably small alpha/beta value for each type of noise distribution. Under differential privacy method, our simulation results show the value of ϵ values corresponding to different generated data set. Our simulation is done for the simple setting of one variable in order to minimize the complexities of data generation. We could assess if it could be used for the more complex and larger data set in the future research. Another important thing is to find what values of ϵ will generate a reasonable statistical disclosure risk in a particular data.

For data swapping method, increasing the number of swapping cells will probably result in loss of data utility. Also, it will take longer time to swapping a great a mount of cells if our sample size is too large. The advantage is that the procedure is simple and programming is very straight-forward. Another disadvantage is that it may distort the correlation between variables and loose its analytic value. For synthetic data method, if there are tons of variables included in the actual data, it will not be a easy task to capture all the relationships in the data. Therefore, our generated data might not contain enough information compared to actual data set and the results from the generated data might misleading as well. The advantage is that this method itself can bear a high risk of disclosure risk and better protect our data.

There are several advantages using noise perturbation method. One attractive feature of addictive noise, for positive quantitative variables, is that it provides uni-

form record level protection to all cells in the data. Second, it's a far simpler and less time consuming procedure, since computer programs for adding noise are much easier to write and modify. Third, all cells are shown in the data and noise can be chosen from any of several types of distribution. However, one concern may be raised that the quality of data and the amount of protection after noise has been introduced.

In our future research, it would be good to explore the features of multiplicative noise and how they perform compared to addictive noise. It would be also good to assess the disclosure risk when we take intruder's prior information into account. The results in our study assume that the intruder does not have any information about how we modify the data. Furthermore, we could evaluate the performances when we apply noise factors from different distributions to different variables in one data set.

Bibliography

Abowd, J. M. and Vilhuber, L. (2008). How protective are synthetic data? In *Proceedings of the UNESCO Chair in data privacy international conference on Privacy in Statistical Databases*, PSD '08, pages 239–246, Berlin, Heidelberg. Springer-Verlag.

Dalenius, T. and Reiss, S. P. (1978). Data-swapping: A technique for disclosure control (extended abstract). In *Proceedings of the Section on Survey Research Methods*, pages 191–194, Washington, DC. American Statistical Association.

Dwork, C. (2008). Differential privacy: A survey of results. In Agrawal, M., Du, D., Duan, Z., and Li, A., editors, *Theory and Applications of Models of Computation*, volume 4978 of *Lecture Notes in Computer Science*, pages 1–19. Springer Berlin Heidelberg.

Elmer, P. J., Obarzanek, E., Vollmer, W. M., Simons-Morton, D., Stevens, V. J., Young, D. R., Lin, P.-H., Champagne, C., Harsha, D. W., Svetkey, L. P., Ard, J., Brantley, P. J., Proschan, M. A., Erlinger, T. P., and Appel, L. J. (2006). Effects

of comprehensive lifestyle modification on diet, weight, physical fitness, and blood pressure control: 18-month results of a randomized trial. *Annals of Internal Medicine* **144,** 485–495.

Fienberg, S. E., Steele, R. J., and Makov, U. (1996). Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: Data swapping and log-linear models. In *Proceedings of the Bureau of the Census*, pages 87–105, Washington, DC: U.S. Bureau of the Census.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., and De Wolf, P.-P. (2012). *Statistical disclosure control*. John Wiley & Sons.

Karr, A. F. and Reiter, J. P. (2014). *Confidentiality and Data Access in the Use of Big Data: Theory and Practical Approaches*, chapter Analytical frameworks for data release: A statistical view, page to appear. Cambridge University Press.

Louis, T. A. (2013). Johns hopkins biostatistics and census bureau. Research @ Census.

Massell, P. B. and Funk, J. M. (2007a). Protecting the confidentiality of tables by adding noise to the underlying microdata. In *Proceedings of the 2007 Third International Conference on Establishment Surveys (ICES-III)*, Montreal, Canada.

BIBLIOGRAPHY

Massell, P. B. and Funk, J. M. (2007b). Recent developments in the use of noise for protecting magnitude data tables: Balancing to improve data quality and rounding that preserves protection. In *Proceedings of the Research Conference of the Federal Committee on Statistical Methodology*, Arlington, Virginia.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* **9,** 538–558.

Nayak, T. K., Sinha, B., and Zayatz, L. (2010). Statistical properties of multiplicative noise masking for confidentiality protection.

Oganian, A. and Domingo-Ferrer, J. (2003). A posteriori disclosure risk measure for tabular data based on conditional entropy. *SORT* **2,** 175.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2014). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-115.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29,** 181–189.

Rubin, D. B. (1993). Satisfying confidentiality constraints through the use of synthetic multiply-imputed microdata. *Journal of Official Statistics* **91,** 461–468.

Schlörer, J. (1981). Security of statistical databases: Multidimensional transformation. *ACM Trans. Database Syst.* **6,** 95–112.

Vita

Lu Shen received the B.S. degree in Statistics and Mathematics from Iowa State University in Aug 2012, and enrolled in the master program at Johns Hopkins University in Sept 2012. She was inducted into the Pi Mu Epsilon Honor Society in 2009, won International Incentive Scholarship in 2010 and 2011, and obtained the Magna Cum Laude distinction in 2012. During her undergraduate study, she made it to the Dean's list consecutively every semester and graduated with honor degree.

In 2014, Lu published two papers including "MicroRNA-Target Binding Structures Mimic MicroRNA Duplex Structures in Humans" and "Genome-wide analysis of regulation of gene expression and H3K9me2 distribution by JIL-1 kinase mediated histone H3S10 phosphorylation in Drosophila. In both papers, she is responsible for all the data anlysis part. She also contributes to the paper "Pre-diabetes and incidence of subclinical myocardial damage" at the Johns Hopkins University and it's in press.